Influence of the Source Content and Encoding Configuration on the Perceived Quality for Scalable Video Coding

Yohann Pitrey[a], Marcus Barkowsky[b], Romuald Pépion[b], Patrick Le Callet[b], Helmut Hlavacs[a]

[a] University of Vienna, Austria

[b] LUNAM Universite de Nantes, IRCCyN UMR CNRS 6597, Nantes, France

ABSTRACT

In video coding, it is commonly accepted that the encoding parameters such as the quantization step-size have an influence on the perceived quality. It is also sometimes accepted that using given encoding parameters, the perceived quality does not change significantly according to the encoded source content. In this paper, we present the outcomes of two video subjective quality assessment experiments in the context of Scalable Video Coding. We encoded a large set of video sequences under a group of constant quality scenarios based on two spatially scalable layers. One first experiment explores of the relation between a wide range of quantization parameters for each layer and the perceived quality, while the second experiment uses a subset of the encoding scenarios on a large number of video sequences. The two experiments are aligned on a common scale using a set of shared processed video sequences, resulting in a database containing the subjective scores for 60 different sources combined with 20 SVC scenarios. We propose a detailed analysis of the experimental results of the two experiments, bringing a clear insight of the relation between the encoding parameters combination of the scalable layers and the perceived quality, as well as spreading light on the differences in terms of quality depending on the encoded source content. As an endeavour to analyse these differences, we propose a classification of the sources with regards to their relative behaviour when compared to the average of other source contents. We use this classification to identify potential factors to explain the differences between source contents.

Keywords: source content, video quality assessment, scalable video coding

1. INTRODUCTION

It is well known that the impact of video coding distortions on the perceived quality is strongly linked with the quantization parameter (QP) used during the encoding process.[1] When Scalable Video Coding (SVC) is considered, the quality of higher layers is determined by the combination of all QP values used for the lower layers.[2] Therefore, studying the performance of different combinations might lead to recommendations in order to optimize the perceived quality.[3] In the video coding community, it is sometimes assumed that under the same encoder configuration, the perceived quality does not vary significantly depending on the source content. However, psychovisual phenomenons are at stake that influence the human judgement of quality depending on the source characteristics under the same encoding configuration.[4] Therefore, some work has been conducted on characterizing the impact of the source content on the visual quality, by using various indicators.[5]

There are two main contributions in this work. First, we present two subjective experiments conducted in the context of SVC, under normalized conditions in order to provide reliable results. A detailed analysis of the results is provided, allowing us to evaluate the impact of 24 MPEG-4 AVC and SVC scenarios on the perceived quality, combined with the influence of 60 different source contents. The processed video sequences as well as the outcomes of the two experiments are made freely available for the research community. As a second contribution, we propose a classification of the 60 source contents with regard to their relative quality when compared to the average over all contents. We present the difficulties of linking this classification with commonly used sequence-level source descriptors and provide a qualitative analysis as an endeavor to orient future work regarding modeling the influence of the source content on the perceived quality.

To contact the authors : [a] firstname.lastname@univie.ac.at [b] firstname.lastname@univ-nantes.fr

2. EXPERIMENTAL DESIGN AND TEST CONDITIONS

We designed two subjective experiments in the context of Scalable Video Coding. The first experiment focuses on the exploration of a wide range of encoding scenarios in terms of quality using spatial scalability. The second experiment extends the first one to a large number of video sources. The design of these two experiments is described in this section, together with the test conditions presented to the observers.

2.1 SVC coding distortions

In the first experiment, 11 source videos were encoded using a wide range of Scalable Video Coding scenarios. The sequences cover several genres such as documentary, sports, outdoors and typical news report clips. They have different levels of spatial and temporal complexities including camera motion, complex scene composition and scene motion. Each video has a duration of 10 seconds, during which no scene cuts appear. The reference videos were extracted from high definition video sequences using cropping and downscaling. First, the videos were cropped from 16:9 to 4:3 to meet the aspect ratio of the VGA format. The cropping step was checked so that the main subject of interest in each sequence remained conveniently framed in the VGA video. Then, the videos were downscaled to VGA format (640×480 pixels) using a freely available downscaling utility.[6]

The 11 source sequences were encoded using two spatial scalable layers, respectively the base layer (L0) in QVGA resolution (320×240) and the enhancement layer (L1) in VGA resolution (640×480), both showing 25 frames per second. A set of 24 constant QP scenarios was designed to cover a wide range of quality levels. We encoded each video using 16 scenarios in which the QP values for the base layer $QP0$ and for the enhancement layer $QP1$ are varied using the values $\{26, 32, 38, 44\}$. Four single layer scenarios using MPEG-4 AVC/H.264 in VGA resolution were also included in the experiment for comparison, using the same QP values. The JSVM reference encoder[6] was used to generate the SVC videos, while the JM reference encoder[7] was used to encode the AVC videos. Both encoders were configured with equivalent parameters. The Group of Pictures size was set to 16 frames, while an I frame was included every 64 frames. Inter-layer prediction was set to the adaptive mode for the JSVM encoder. After encoding, the videos were decoded using their respective reference decoding software. Additionally, four single layer scenarios in QVGA were added to the experiment. The videos were first encoded using the JM reference encoder, then decoded and upscaled to VGA resolution using the Lanczos upscaler found in the JSVM software suite. The unprocessed VGA videos were included in the experiment as hidden references. Finally, five scenarios including coding and transmission errors were included, as a mean to align this experiment on a common quality scale with other sets of data, following the method presented in our previous work.[8] These conditions serve as a common set of videos between several experiments and are not the main focus of the presented experiment. A similar alignment process is applied between the two current experiments and will be discussed in section 3.2. In total, 330 processed video sequences (PVS) were included in the first experiment.

2.2 Influence of the source content under SVC distortions

For the second experiment, we encoded a total of 60 video source sequences using a subset of the 16 SVC and the 4 upscaled AVC configurations from the first experiment for each sequence. The 60 sequences were selected from freely available video databases and were all either available in VGA at 25 frames per second, or were downscaled to this format using the same procedure as described for the first experiment. The sequences contain a wide variety of genres, such as natural and rendered scenes, documentary, news reports and sports clips. Some of them were extracted from professionally produced contents and feature noticeable editing choices such as slow motion, aesthetic scene composition and dramatization. Some sequences include scene cuts or significant camera motion such as panning and tilting. Therefore the level of complexity and activity in each video varies on a large scale, such as illustrated by the Spatial and Temporal Indicators[9] displayed on Figure 1.

Naturally, showing the 60 video sequences encoded with all the scenarios in a single subjective experiment would not allow us to respect the constraints on the duration of an evaluation session, besides representing a prohibitive amount of efforts in testing the full range of configurations. As a result, we designed the experiment so that each video sequence is displayed under only five different encoding conditions. One of these five conditions was the non-coded reference, to follow the standard recommendations regarding high quality anchors. Then, we

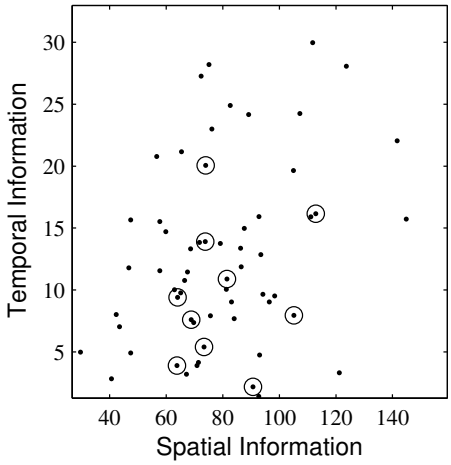

Figure 1. Average spatial/temporal indicators for the 60 source contents. The common sources for the two experiments are marked with circles around the dots.

selected four conditions expected to cover the full quality scale for each sequence in a semi-random way. The selection process being one important particularity of this experiment, it is here described in details.

As we conducted the two experiments several monthes apart from each other, we had access to the results of the first test when designing the second one. We first ordered the conditions from the first test in terms of decreasing MOS. As we used a 5-category scale, the MOS values were comprised between 5 (excellent quality) and 1 (bad quality) (the rating conditions are fully described in section 2.3). The conditions were split into four groups: *Group 1:* conditions with a MOS comprised between 4 and 5, *Group 2:* conditions with a MOS comprised between 3 and 4, *Group 3:* MOS between 2 and 3, and *Group 4:* MOS between 1 and 2. After building the groups, one condition was selected randomly from each group, for each of the 60 source sequences. As a result, each video content was encoded with four levels of quality covering the full scale, while two given contents did not share the same four conditions. This allows us to distribute the 20 encoding scenarios equally among the 60 contents. A total of 300 PVS were included in the second experiment. The next section describes the test conditions, which were equivalent for the two experiments.

2.3 Test conditions

The test conditions were set according to the ITU-R BT.500 recommendation for the room setup, correct illumination and display calibration. The sequences were displayed without upscaling on a 40 inch TV-Logic LMV401 reference LCD display operating at its native resolution of 1920×1080 pixels at 60Hz, placed at a distance of 4 times the height of the displayed videos from the observer. As the VGA sequences did not cover the whole HD screen, a large gray border with a pixel value of Y=108 in the YCbCr color space was placed around them. This value corresponded to 25% of the maximum brightness of the display.

We used an Absolute Category Rating with Hidden Reference (ACR-HR) protocol for the observers to rate the videos, such as described in the ITU-T P.910 recommendation. The display order of the videos was randomized following the constraints from the VQEG Multimedia testplan, ensuring that two successive PVSes were not from the same source sequence. We used a 5-level category rating scale, with the classical labels ranging from "excellent" to "bad".

A total of 27 non expert viewers participated in the first experiment. Their age range was 18 to 61 years old, with an average of 29.25. They were tested for visual acuity using a Snellen chart and for color blindness using Ishihara plates. Two sessions of 25 minutes were conducted with each observer, separated by a 10 to 15 minutes break. Before the real experiment, the observers were presented a short training session. For the second experiment, 36 observers took the test (age range: 18–46, average: 26.16), after the same verifications and under the same conditions as for the first experiment.

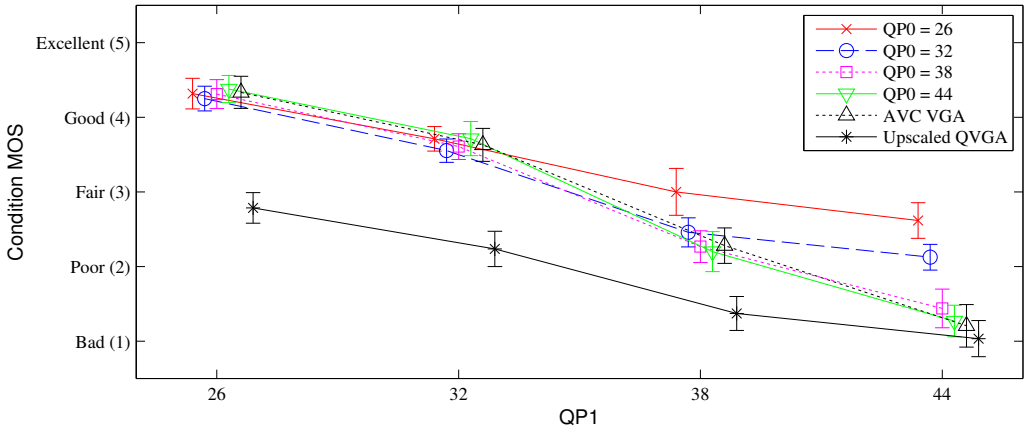

Figure 2. Condition MOS for the SVC coding distortions experiment depending on the values of QP0 and QP1. The 95% confidence intervals are displayed as vertical error bars. A slight jitter was added to the values of the x-axis coordinates to ease reading the plot.

3. EXPERIMENTAL RESULTS

The test PVS and the observer ratings for the two experiments are made freely available for the research community to use. More details are available at the address provided in [10].

3.1 SVC coding distortions

Figure 2 displays the condition MOS obtained for each encoding scenario as a function of the QP values for the two layers. One can observe that the QP of the highest layer mostly determines the final MOS. When QP1 is equal to 26 and 32, the four combinations of QP0 obtain very similar MOS values. This result also applies when the value of QP0 is equal to or higher than the value of QP1. It is interesting to mention the consequences of this observation on the encoded bitrate. Encoding the base layer with high quality does not only have no impact on the overall quality, but also increases the output bitrate needed to encode the video stream. However, a high quality base layer can represent a significant interest in the case of transmission errors, as it can be used to perform error concealment.[11] A different tendency appears when QP0 is lower than QP1. One can notice that when QP0 is equal to 26, the two scenarios with QP1 equal to 38 and 44 reach equivalent quality, based on the significant overlapping of their intervals of confidence. This illustrates that the previously observed dominance of the higher layer is inverted, and the quality of the base layer mostly determines the overall quality.

Regarding the comparison between MPEG-4 AVC and SVC, one can observe that the four AVC VGA scenarios are equivalent in terms of MOS to the four SVC scenarios where QP0 is equal to QP1. Naturally, an increase in bitrate was observed for the SVC configurations, as two layers have to be encoded. This increase, also involving extra computation needs, is hard to justify in a coding distortion-only context, but can again be advantageous in case of transmission errors. Some interesting results are to be observed for the upscaled QVGA scenarios. When the QP value is equal to 26 and 32, the quality obtained by the upscaled videos is equivalent to the quality of the SVC and AVC scnearios encoded with QP1 equal or superior to 38 (except for QP0 = 26, for which the quality is higher). In this case, encoding only one layer with reduced resolution and a relatively high quality is as good as encoding either one full resolution layer or even two scalable layers with lower qualities.

3.2 Aligning both experiments on a common scale

The two presented experiments complement each other. The first experiment provides subjective outcomes on a wide set of SVC configurations yet on a limited number of video contents, while the second provides outcomes on a large number of video contents, yet on a limited number of encoding scenarios. As a result, comparing the outcomes of the two experiments to each other would allow us to estimate the behaviour of the 60 sequences from the second experiment on the 20 encoding scenarios from the first experiment.

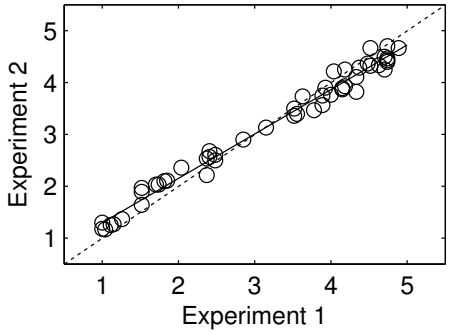

Figure 3. Scatter plot showing the MOS obtained by the PVS of the common set between the two experiments. we modeled the relation between the two scales by using a linear function: $MOS_{exp1} = 0.8577 * MOS_{exp2} + 0.4355$.

However, in order to compare the results from the two experiments to each other, their outcomes first need to be aligned on a common scale. This procedure helps compensating the context effect, following which observers tend to use the full quality scale for a given experiment, making each experiment a specific environment with a unique mapping between the characteristics of a processed video and the observers ratings. In our previous work, we presented an approach aimed at providing a low-cost yet efficient way to align several experiments on a common scale.[8] This approach is based on a common set of PVS included in each experiment. Under the constraint of fair repartition of the scores on the quality scale, this common set can be used as a way to determine a mapping function from the outcomes of one experiment onto the scale of the second experiment. A total of 50 PVS were shared by the two experiments. Figure 3 presents the scores obtained by these PVS in each experiment. As one can easily observe, the relation between the scores obtained by the common PVS in the two experiments can be modeled using a linear function. We used linear regression to find the coefficients of this function and map the outcomes of the first experiment on the scale of the second experiment.

After the alignment process, we consider that the two experiments share the same quality scale, and therefore compare the results to each other. We use the data from the first experiment as a ground truth representing a comprehensive description of the relation between encoding parameters and subjective quality. Then, we analyse the relative quality of the 60 source sequences from the second experiment when compared to this ground truth.

3.3 Influence of the source content on perceived quality

Figure 4 displays the comparison of the condition MOS and intervals of confidence from the first experiment to the individual qualities reached by the 60 source contents from the second experiment. One can observe a significant variation of the MOS according to the source content, reaching approximately one MOS category in the most extreme cases. Due to the semi-random design of this experiment, a direct comparison of different contents is not possible, as two contents do not share the same four encoding configurations. However, the encoding configurations for each video sequence were selected from groups with comparable quality levels. As a result, we propose to compare the behaviour of the different contents in regard to the five groups described in section 2.2.

A particularly interesting result when observing the behaviour of the individual source contents in the five groups of quality is that the Pearson linear correlation between the five conditions and the average over the 11 contents from the first experiment over the same encoding scenarios is very high (average = 0.9924, standard deviation = 0.0077). This means that there is a close relation between the condition MOS observed on the ground truth from the first experiment and the behaviour of each source content from the second experiment. One has to notice that the correlation is obtained on five data points (*i.e.*: the number of available points for each content in the second experiment). Therefore, with only three degrees of freedom, its reliability could be discussed. However, considering the very high values observed on all 60 contents, we consider that this is a significant result.

A consequence of this high correlation is that the behaviour of a given source content over the 20 SVC scenarios could be approximated from the 5 semi-randomly selected scenarios in the second experiment. We

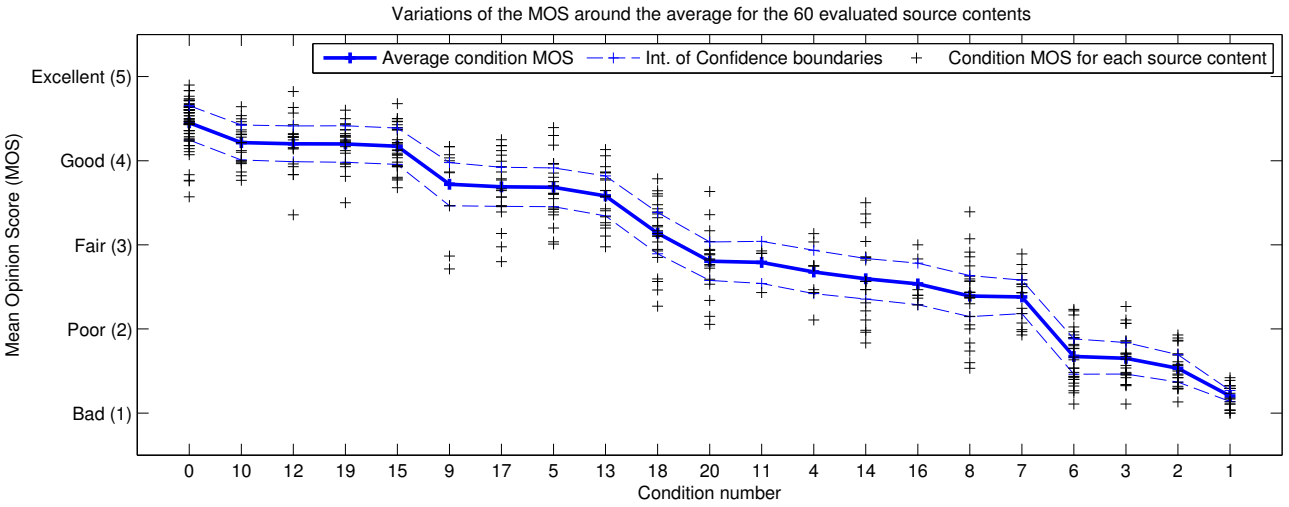

Figure 4. Variability among 60 video sequences when compared to the condition MOS under a wide range of qualities.

propose to use this result to analyse the relative behaviour of each content when compared to the average quality from the first experiment. To this end, we analysed the statistical significance of the difference between the quality obtained for a given content under a given encoding scenario and the average quality obtained for this scenario on the first experiment. The statistical significance was calculated by means of a paired Student t-test between the two variables. The outcomes of this test were then discriminated using three categories: *"the PVS MOS of content S under condition H is significantly higher than the MOS of condition H"*, *"no significant difference can be observed between the PVS MOS of content S under condition H and the MOS of condition H"* and *"the PVS MOS of content S under condition H is significantly lower than the MOS of condition H"*. This simplification of the problem was motivated by the fact that the maximal MOS variations among all 60 contents from the average are approximately contained within one MOS category on the 5-level scale, such as illustrated on Figure 4. Hence, with the indication on the variability of observers votes given by the size of the confidence intervals, the three proposed categories allow for a meaningful level of comparison.

Based on these three categories, we performed a detailed analysis of the behaviour of the source contents relatively to the average. This analysis revealed that several groups could be clearly identified among the 60 contents. Figure 5 illustrates the main tendencies of five categories of contents. A first category (Cat. A on Fig. 5) groups contents that reach significantly higher quality than the average, consistently over their five PVS. Category B groups the contents for which no significant difference with the average can be identified. Category C groups the contents obtaining a significantly lower quality than the average. Categories D and E are intermediate categories. Category D covers contents for which the quality decreases faster than the average when considering the condition MOS in decreasing order, while category E covers the contents for which the quality decreases slower than the average. The number of contents classifed in each category were 17, 17, 6, 15 and 5, respectively from Cat. A to E.

As an endeavour to identify indicators explaining these different contents categories, we calculated the values of several video source descriptors on the 60 contents and injected them in the WEKA data mining framework.[12] First, we processed the average values of the Spatial and Temporal Information descriptors,[9] which are often used as simple measures to analyse a video in terms of coding complexity. Then, we used the MSU Video Quality Measurement professional tool[13] to process the levels of *blocking*, *blurring*, *noise* and *brightness flicking* on all the sequences. Only the non-coded reference video sequences were used to process these indicators, as an attempt to use only the characteristics of the content itself to predict its behaviour under various levels of coding distortions. We used the implementation of the Support Vector Machines classifier from WEKA with standard parameters and a 10-fold cross-validation procedure to classify the 60 sequences using the listed descriptors. This resulted in a percentage of correct classifications as low as 31.6% over the whole set of source contents. Several reasons can explain this low classification accuracy. First, the proposed content categories (*i.e.*: A to

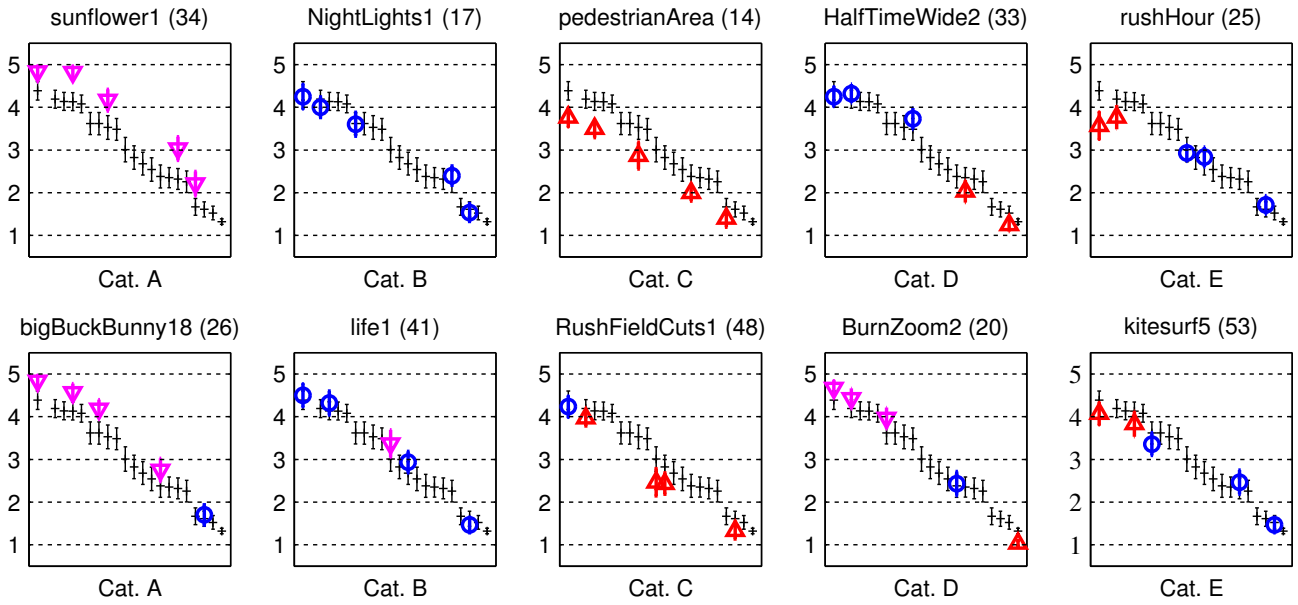

Figure 5. Example of relative behaviour of 10 source contents when compared to the adjusted average of condition MOS values. Configurations for which the quality is significantly lower than the average are identified by an upward-pointing triangle "△"; configurations for which no significant difference can be identified are identified by a circle "◯"; configurations for which the quality is significantly higher than the average are identified by a down-pointing triangle "▽". The confidence intervals are displayed as vertical error bars to give an indication of the variability of votes among observers.

E) based on our observations could be unadapted for content classification. We reduced the dimensionality of the problem by introducing three simplification steps, *i.e.*: quality level category, statistical significance of the difference with the average and global tendency of the relative quality. This simplification allowed us to take the variability of the votes among observers into account and avoid trying to predict the MOS beyond the inherent variability of the problem. A second reason to explain the low classification accuracy is that the calculated indicators are not sufficient to match the source characteristics to the observed categories. We decided to use sequence-level indicators for their simplicity and the fact that they tend to be commonly used to describe video sequences in practical applications. Our intuition was that these indicators are indeed not enough to predict the behaviour of a given content under different levels of encoding quality. Last, a third possible explanation could be that considering only the non-coded reference video does not ease describing its behaviour. Using the proposed indicators as full reference measures on the coded videos could allow for a better understanding of the behaviour of the source contents under coding distortions.

Despite the low accuracy of the attempted classification, a qualitative analysis of the contents with regard to their category reveals some common characteristics. Category A contains the majority of rendered contents included in the experiment. These contents reflect particularly high editing efforts, in terms of colours, contrast, scene composition and camera motion. A significant amount of sequences from Category A share such characteristics (*e.g.*: water-sports clips picturing slow motion). Most sequences in this category convey clear messages or picture visually or dramatically appealing content. Sharpness of the pictured objects, as well as rich colours and good contrast are also common elements. Close-up and macro shots are used in a significant amount of videos. Category B is difficult to characterize clearly. Some of the contents picture appealing content, but this advantage is decreased by the presence of unsufficient lighting or contrast, complex moving structures such as water or vegetation in wind. For a significant amount of contents from this category, it is difficult to determine the context in which the video was produced, or to identify a clear message from the 10 seconds of playback. In Category C, the recurrent features are motion blur (*e.g.*: action shots or complex continuous motion), the presence of small structures due to wide angle shots, insufficient lighting in the regions of interest. The absence of clear message also appears to be a common element to the sequences of Cat. C. Categories D and E are

difficult to describe because of their intermediate nature. Category D contains sequences for which the higher quality encoding scenarios show an advantage when compared to the average, while the low quality scenarios show a disadvantage. From our observations, this behaviour could be related to the presence of elements that easily suffer from coding distortions or are difficult to encode (*e.g.*: small structures, non rigid motion, light flashes, high motion). Finally, Category E contains sequences for which the non-coded reference was rated with lower quality than the average, whereas the coded scenarios exhibit equivalent or higher quality than the average. We observed that most of the sequences in this category contain elements that can be mistaken for coding distortions, including in the non-coded reference sequences (*e.g.*: hot-air smoke effect, oscillating leaves in the wind with granular texture, laboured earth pattern). The main type of distortion in the experiment being related to coding, the observers might have been influenced to identify such effects as negative for the quality.

4. CONCLUSION

Two subjective experiments exploring the influence of the source content under a wide range of quality levels in the context of Scalable Video Coding were presented in this paper. We aligned the two experiments on a common scale using a set of common videos, and used the outcomes of the first experiment as a ground truth to evaluate the relative behaviour of each source content under five levels of quality. After reducing the problem to a limited number of content categories, we calculated common source descriptors to attempt to reproduce the observed categories. The classification exhibited a low accuracy, which brought us to identify several issues questionning the possibility to describe the behaviour of a particular content using simple sequence level descriptors on the non-coded reference. A qualitative analysis was then presented as an attempt to identify higher-level parameters having an influence on the relative quality, granting an advantage to the visually appealing and well-edited contents, while penalizing poor lighting or contrast and sequences that do not deliver a clear message.

ACKNOWLEDGMENTS

The authors would like to thank Z. Shahid for his contribution in the design of the second experiment presented in this paper. The work presented in this paper was supported by the SVC4QoE project funded by the French DGE (http://dot-projets.degetel.com/SVC4QoE/), and the CACMTV project funded by the Viennese WWTF Research Funding Organization (http://www.ani.univie.ac.at/˜cacmtv)

REFERENCES

[1] Wolff, T., Ho, H.-H., Foley, J. M., and Mitra, S. K., "Modeling subjectively perceived annoyance of H.264/AVC video as a function of perceived artifact strength," *Signal Processing* **90**, 80–92 (Jan. 2010).
[2] Lee, J.-s., De Simone, F., and Ebrahimi, T., "Subjective Quality Evaluation via Paired Comparison: Application to Scalable Video Coding," *Multimedia, IEEE Transactions on* **6**(99), 1–1 (2011).
[3] Unterweger, A. and Thoma, H., "The Influence of Bit Rate Allocation to Scalability Layers on Video Quality in H. 264 SVC," in [*Picture Coding Symposium (PCS2007)*], **1** (2007).
[4] Mantel, C., Kunlin, T., and Ladret, P., "The role of temporal aspects for quality assessment," in [*Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*], 94–99, IEEE (2010).
[5] Moorthy, A., Obrador, P., and Oliver, N., "Towards computational models of the visual aesthetic appeal of consumer videos," *Computer VisionECCV 2010* , 1–14 (2010).
[6] Joint Video Team, "JSVM Reference Software, Version 9.18." http://ip.hhi.de/imagecom_G1/savce/downloads/.
[7] Joint Video Team, "Joint Video Model Reference Software." http://iphome.hhi.de/suehring/tml/.
[8] Pitrey, Y., Engelke, U., Barkowsky, M., Pépion, R., and Le Callet, P., "Aligning subjective tests using a low cost common set," in [*Workshop on Quality of Exp. for Multimedia Content Sharing (QoEMCS) - Euro ITV*], (2011).
[9] Pinson, M. and Wolf, S., "A New Standardized Method for Objectively Measuring Video Quality," *IEEE Transactions on Broadcasting* **50**, 312–322 (Sept. 2004).
[10] IRCCyN IVC research group freely available video subjective databases. http://www.irccyn.ec-nantes.fr/spip.php?article491.
[11] Pitrey, Y., Barkowsky, M., Callet, P. L., and Pépion, R., "Evaluation of MPEG4-SVC for QoE protection in the context of transmission errors," *SPIE Optical Engineering* (2010).
[12] The University of Waikato, "WEKA Data Mining Software." http://www.cs.waikato.ac.nz/ml/weka/.
[13] V. Yoockin, A. Ratushnyak, "MSU Video Quality Measurement Tool Professional Version 2.7.3." http://compression.ru/video/quality_measure/video_measurement_tool_en.html.